Solving the problem of loan default problem in the banking sector using machine learning

Deepank Shokeen , Vibhor Grover and Vansh Verma

Netaji Subhas University of Technology

ABSTRACT

Loans are a really elementary supply of any bank's revenue, in order that they work indefatigably to create certain that they solely provide loans to customers who won't neglect the monthly payments. A lot of attention is paid to the current issue and use varied ways in which to notice and predict the default behaviours of their customers. However, a lot of the time, due to human error, they'll fail to envision some key info. This paper proposes an improved approach exploitation of machine learning approaches like KNN, decision tree, random forest, Xgboost and logistic regression to predict defaulters. This can help banks conserve their workforce and financial resources by reducing the quantity of steps they need to require so as to visualize if someone is eligible for a loan. It turns out that XGBoost and different models show very little distinction ,on basis of the domain understanding and analysing different performance metric XGBoost was the one outperforming other existing classification models.

1.Introduction

Loan disposition plays a crucial role in our way of life and powerfully promotes the expansion of consumption and therefore the economy [1]. Taking a loan has been inevitable for individuals since individuals round the world rely on loans to beat monetary constraints to realize their personal goals, and organizations trust loans to expand their production. In most cases, loan lending is helpful to each the borrowers and the lenders. However, loan default remains unavoidable, that carries a good risk and should even find yourself in an exceedingly financial crisis. Therefore, it's notably important to spot whether or not a candidate is eligible for receiving a loan. The increasing range of dangerous debts ensuing from industrial banks' loans reflects the growing drawback of agitated banks within the economic system. We've got used data processing algorithms to predict the possible defaulters from a dataset that contains info regarding consumer credit applications, thereby serving to the banks for creating higher choices within the future.

In the past, the analysis primarily trusted manual review ,that was long and labour-intensive. Recently, banks have opted for machine learning approaches to mechanically predict the loan default since it will extremely enhance the accuracy and also the potency of the prediction [2]. On the one hand, banks can collect a vast quantity of group action information because of the prosperity of on-line searching and mobile payments. On the opposite hand, machine learning models are chop-chop evolving and have winning applications in varied fields, motivating the bank trade to use them to predict loan default.

Using various pre-processing data mining steps, generating new metrics and combining with a classification algorithm we were able to come up with a well performing outcome. We check our framework with 5 of the foremost used classification models within the credit risk literature: Logistic Regression [3], Decision Tree [4], Random Forest [5], Extreme gradient boosting (XG Boost)[6], KNN[7] by evolving new metrics. Employing a public information out there on Kaggle.com for default

prediction, we have a tendency to reason the potential model risks from confirming these models and their potential prognostic performance.

2. Related work

In this section, we will discuss some of the existing works on loan default prediction.

For default prediction, authors of [8] propose more efficient strategies using machine learning methods such as ANNs, decision trees, SVMs, and logistic regression. Metrics such as log loss, Jaccard similarity factor, and F1 score to measure the accuracy of these approaches were proposed. Compare these measures to evaluate the accuracy of your predictions. By minimizing the process required to determine whether a loan is eligible, banks can save human and financial resources.

The introduction of new machine learning (ML) algorithms for predicting loan defaults is associated with better predictive performance [9]. However, it also creates new model risks, especially with respect to the supervisory review process. Recent industry research frequently notes that uncertainty about how regulators can assess these risks can be a barrier to innovation. In the study, authors of [9] propose a novel framework for quantifying model risk adjustments to compare the performance of different machine learning methods.

The aims to classify whether people are able to pay off their debt while preventing banks from incurring significant losses by the authors of [10]. A defaulter could bankrupt a bank by failing to pay large loans, which could lead to a financial crisis for the country or any bank providing credit. As the number of borrowers increases, so does the number of people who do not repay their debts. There are many classification machine learning techniques and deep learning approaches that can be used to solve problems. The main purpose of the study was to compare and contrast Random Forest, Logistic Regression, and XG Boost models to determine which model provides the most accurate performance.

As the demand for bank loans increased, the possibility of non-performing loans, or defaults, also increased significantly [11]. Authors of [11] develop machine learning algorithms to solve problems that can reduce credit risk and improve service efficiency, especially when faced with data imbalance problems. First, training of random forest model using historical bank loan data and relevant data from other financial institutions was done. Secondly, modification of the algorithm for classifying unbalanced data with random forest and fine-tuned data feature extraction methods was done. Thirdly, the results showed that the machine learning risk prediction algorithm outperformed the traditional statistical algorithm. In addition, random forest algorithms was used to determine the impact of data characteristics, allowing for more accurate assessment of credit risk in the financial sector by obtaining characteristics that have a large impact on outcome decisions.

The two most important questions in the banking industry have been: (i) How risky is the borrower? (ii) Given the borrower's risk, should one lend to the borrower?[12]. In light of the issues raised, authors of [12] proposes two machine learning models to assist banking authorities by evaluating specific attributes to predict whether they should get a loan and facilitating the process of selecting suitable individuals from a given list of applicants for a loan. The article provided a comprehensive and comparative analysis of two algorithms: (I) random forests and (ii) decision trees. Both algorithms were used on the same data set, and conclusions were drawn based on the results that the random forest algorithm outperformed the decision tree algorithm with much higher accuracy.

Another study focused on modelling and predicting the willingness to repay a credit card loan[13]. Methods used in that study include machine learning using random forest approaches, artificial neural

networks, support vector machines, logistic regression, and naive bayes. 11 variables were analysed and the performance of five methods was compared with ROC and AUC scores. The results of the study were. The random forest method was considered the best for processing the basic credit card dataset with an AUC of 89%. That model can contribute to the resolution of possible defaults and is a huge boon to the credit card industry. From the manager's point of view based on the PDP, it can be judged that the higher the income and credit card limit of 7 million to 50 million won, the higher the possibility of default.

Another study analyses the performance of a suite of machine learning models in predicting underlying risk using standard statistical models such as logistic regression as a benchmark[14]. It was seen that machine learning models offer significant advantages in discriminant power and accuracy over statistical models when only a limited set of information is available, such as in the case of external credit risk assessment. This benefit diminishes when confidential information such as credit behaviour indicators are also available and is negligible when the data set is small. It also uses machine learning score-based credit allocation rules to evaluate the impact on total credit supply and the number of borrowers accessing credit. Machine learning models reduce lenders' credit losses by focusing most of their credit on safer, larger borrowers.

3. Methodology

3.1. Dataset Preparation

The dataset used in this research is a real world data publicly available on Kaggle. It includes loan data with 34 columns (features) and 174000 rows. The data was in the form of a CSV file.

Diagram

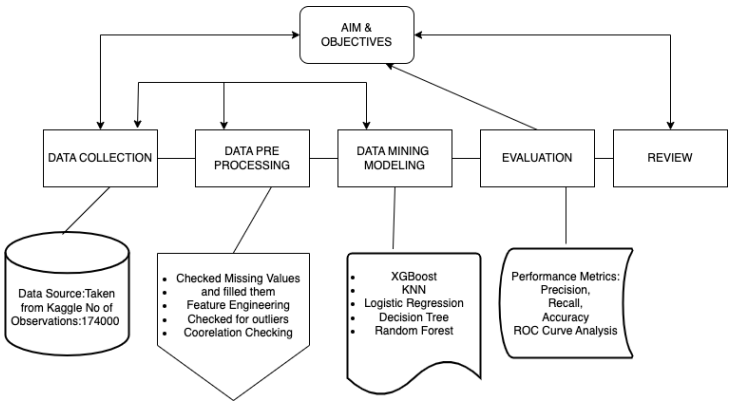

Figure 1. Flowchart of the methodology

In Figure 1. the flowchart explains the complete process of our methodology from data collection to data pre-processing to data mining modelling then evaluation and reviewing our complete process. The details of the following processes is as follows:

3.2.Data Pre processing

Data Pre Processing is the transformation of the raw data into more meaningful and efficient format which would help the machine learning models to perform better [15]. The real world data is generally

full of noisy data,with missing values and duplicate data but for machine learning algorithms to perform efficiently we need the data to be in processed form which is the aim of this step.

We performed the following pre-processing steps on our data.

i. We removed the unimportant features which did not provided any viable information. These features were 'id', 'year', 'loan_limit', 'gender', 'approv_in_adv' , 'credit_worthiness', 'open_credit','business_or_commercial','interest_rate_spread', , 'neg_ammortization' ,'construction_type', 'occupancy_type', 'secured_by', 'submission_of_application', 'ltv',and a few more unnecessary features.
ii. We had to fill the missing values to increase the accuracy of the model so we filled the missing values according to the features. Most of the features missing values were filled by taking the mean of the non-null values and the missing values of the feature 'age' was filled with the mode value.
iii. We had some outlier values for 'income' feature which was found using data visualization these values were then replaced by the mean income value.
iv. We also had outlier with 'interest rate' feature where it was zero which was also then replaced by the mean interest rate values.
v. We had many categorical features in our data, these were converted into numerical data by encoding. These features include 'loan type' and 'age'.
vi. We also introduced a new metric to better utilize the multicollinearity of the features.

$$LRP\ Score = \frac{Loan\ Amount * Property\ Amount}{Rate\ of\ Interest}$$

$$LC\ Score = \frac{Loan\ Amount}{Credit\ Score}$$

vii. Synthetic minority oversampling technique (SMOTE) was done to solve the imbalance in our dataset.
viii. Data partitioning was then done to train our ML model.

Here are the feature after the data pre-processing stage:

Features	Description
Loan_type	Type of loan applied for
Loan_amount	Value of the Loan
Rate_of_Interest	Interest Rate applicable for the loan
Term	Duration of the Loan
Property_value	Collateral property value
Income	Income of the applicant
Credit_score	Credit score of the applicant
Age	Age of the applicant
Status	Current loan status
LRP Score	Combination of Loan amount, property amount and rate of interest
LC Score	Combination of Loan amount and Credit Score
Dtir1	Debt to income ratio

Table 1.Features after Pre-Processing

3.3. Model Training

In our experiment, we would we using XGBoost machine learning algorithm to train our model. The dataset was partitioned into two parts one containing 80% of the entire dataset and other containing 20%. The 80% part is the training part and rest is for testing purpose. We use the training dataset to train the model and used the testing dataset to evaluate our model obtained.

XGBoost stands for Xtreme Gradient Boosting, it is a decision tree based algorithm and also an improvement over the gradient boost algorithm. This technique generalizes the component classification decision trees by utilizing arbitrary differentiable loss function for model optimization.

Steps involved in XGBoost algorithm are:

i. First, we input the training set $\{(xi,)\}$, a loss function $L(y,F(x))$, weak learners *S* and a learning rate α.

$$\hat{p}_{(0)}(x) = \arg min_{\theta} \sum_{i=1}^{N} L(y_i, \theta) \quad (1)$$

ii. Initialize model with a constant

For *s* = 1 to *S*:

iii. Compute the gradient (q) and the hessians (t):

$$\hat{q}_s(x_i) = \left[\frac{\partial L(y_i, p(x_i))}{\partial p(x_i)}\right]_{p(x) = \hat{p}_{s-1}(x)} \quad (2)$$

$$\hat{t}_s(x_i) = \left[\frac{\partial L(y_i, p(x_i))}{\partial p(x_i)^2}\right]_{p(x) = \hat{p}_{s-1}(x)} \quad (3)$$

iv. Fit a base learner using the training set $\left\{x_i, \frac{-\hat{q}_s(x_i)}{\hat{t}_s(x_i)}\right\}$ by solving the optimization question below:

$$\hat{\varphi}_s = \arg min_{\varphi} \sum_{i=1}^{N} {}^{1}/_{2}\, \hat{t}_s(x_i) \left[-\frac{\hat{q}_s(x_i)}{\hat{t}_s(x_i)} - \varphi(x_i)\right]^2 \quad (4)$$

$$\hat{p}_s(x) = \alpha\, \hat{\varphi}_s(x) \quad (5)$$

v. Update the model:

$$\hat{p}_s(x) = \hat{p}_{(s-1)}(x) + \hat{p}_s(x) \quad (6)$$

vi. Output:

$$\hat{p}(x) = \hat{p}_{(x)}(x) \sum_{s=0}^{S} \hat{p}_s(x) \quad (7)$$

3.4.Results And Evaluation

The dataset is enormous & consists of multiple deteministic factors like borrower's income, age, credit score. The dataset is subject to strong multicollinearity & empty values that is why we tried feature selection and event making some new features by combining the existing ones.

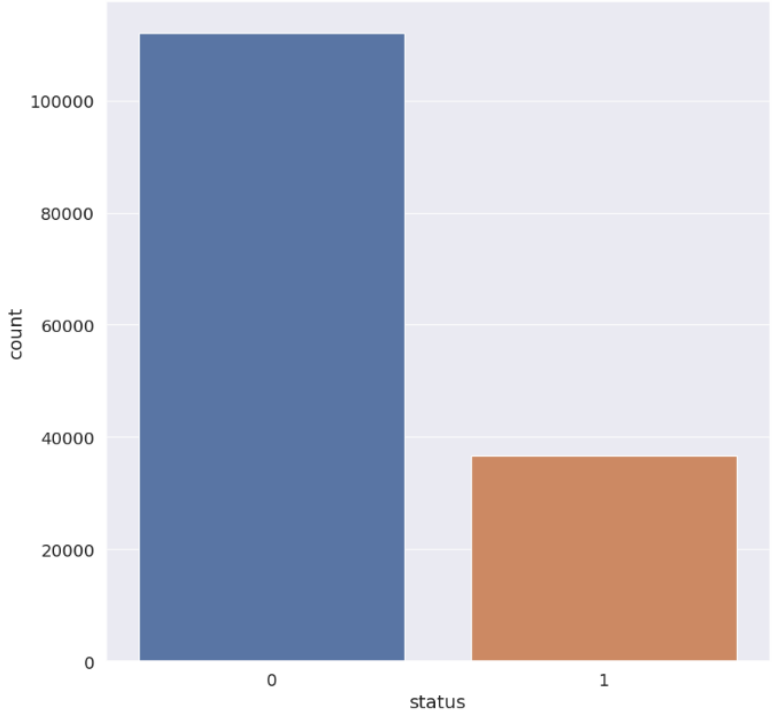

Figure 2. Bar graph of distribution of status

From figure 2. We can observe that our dataset is imbalanced as we have more number of people not defaulting and less number of people defaulting.

We applied various common machine learning algorithms along with our XGBoost algorithm to have a comparative study. We calculated various performance metrics [16], accuracy and ROC Curve to have a detailed looks at the outcomes.

Performance Metrics:

Precision: It is the measure of how many of the positive values are predicted correct.

$$Precison = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (8)$$

True Positive means the sample which are likely to default are classified as likely to default as well.

False Positive means the sample which are not likely to default are classified as likely to default.

Recall: It is the percentage of the true positives that were predicted correctly.

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (9)$$

False Negative means the sample which are likely to default are classified as not likely to default.

Accuracy: It is the percentage of the number of correctly classified data instances over the total number of instances.

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (10)$$

True negative means the sample which are not likely to default are classified as not likely to default as well.

	Accuracy	Precision	Recall
XGBoost	**97.77**	**93.7**	**97.59**
Decision Tree	96.87	93.3	94.24
Random Forest	96.2	91.5	93.4
KNN	80.5	62.99	54
Logistic Regression	49.16	25	51.74

Table2. Performance Metrics of several classification Models

We can observe from Table2 that our algorithm XGBoost is outperforming all the others in terms of the accuracy. Since our problem is of banking loan default which would be an imbalanced dataset with less no of people defaulting and more no of people not defaulting so in such scenarios we prefer Precision and Recall as our main performance metric, and since we would prefer to have less number of false negative because one would not desire to have a person likely to default classified as not defaulting, so Recall is our main evaluation metric and XGBoost had significantly better Recall than others.

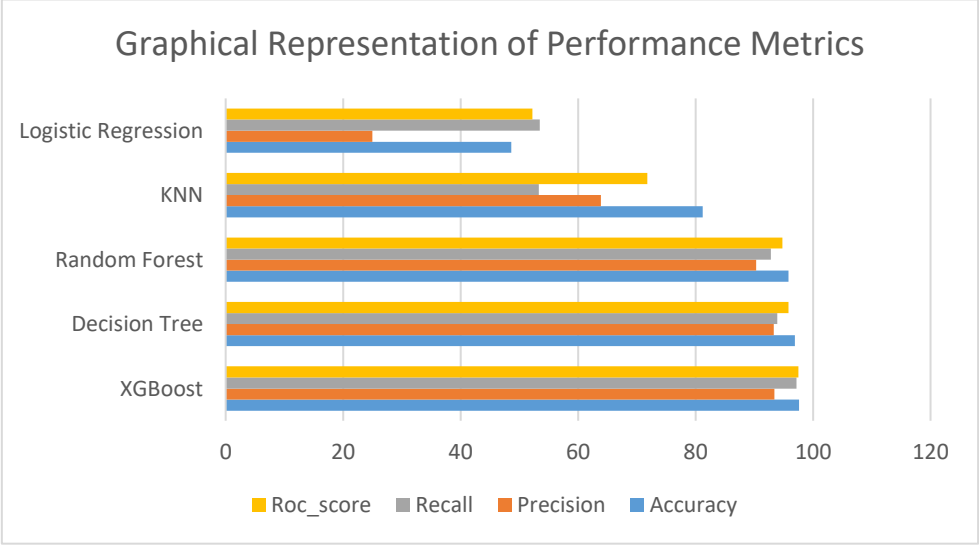

Figure3.Bar Chart of Perform Metrics

In Figure 3 Blue bars represent the Accuracy of all the models, Orange represents the Precision, Grey represents the Recall and yellow bars represents the ROC scores.

ROC Curve: Receiver Operating Characteristics is a plot of the probability of a true positive against the probability of a false positive for all possible threshold values.

Let’s also have a look at a ROC curve including all our ml models as well.

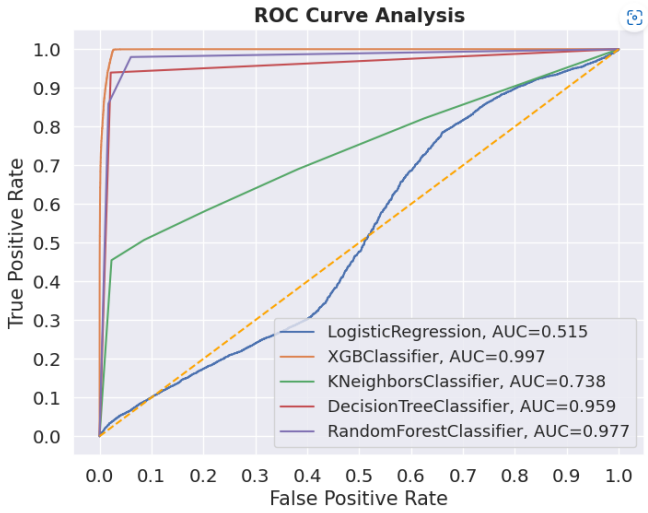

Figure4.ROC curve of all ML models

In Figure 4, the x-axis represents the false positive rate and the y-axis represents True Positive Rate, and the area under the curve represents the measure of usefulness of a test. The orange curve represents XGBoost Model, purple curve represents Random Forest Model, red curve represents Decision tree Model, green curve represents KNN Model and blue curve represents Logistic Regression Model.Logistic Regression model did not perform well for our dataset as clearly we can see for none of the threshold is the True positive rate high and false positive rate low .Xgboost, Random forest and Decision tree performed really well.

From figure 4 we can observe that the XGBoost model’s ROC Curve is closest to the upper left corner and has the largest area under the curve, indicating the best performance as we were able to get high positive rate and less false positive rate.

4. Conclusion

Loan default problem is still one of the most challenging and important problem in the banking industry for their fluent working. In our report, we proposed a solution to the loan default problem using the XGBoost algorithm.

We compared our algorithm to various other machine learning algorithms which included Logistic Regression, Decision Tree Classification, K Nearest Neighbour classification and Random Forest classification and found that our algorithm outperformed all of them. So we can conclude that our algorithm can effectively achieve loan default prediction and reduce the external credit crisis caused by customer default faced by the banking industry.

References

1. Lai, Lili. "Loan default prediction with machine learning techniques." In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pp. 5-9. IEEE, 2020.

2. Wang, Yuelin, Yihan Zhang, Yan Lu, and Xinran Yu. "A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data." *Procedia Computer Science* 174 (2020): 141-149.

3. Sperandei, Sandro. "Understanding logistic regression analysis." *Biochemia medica* 24, no. 1 (2014): 12-18.

4. Hauska, Hans, and Philip H. Swain. "The decision tree classifier: design and potential." In *LARS Symposia*, p. 45. 1975.

5. Ho, Tin Kam. "The random subspace method for constructing decision forests." *IEEE transactions on pattern analysis and machine intelligence* 20, no. 8 (1998): 832-844.

6. C Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.

7. Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "KNN model-based approach in classification." In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 986-996. Springer, Berlin, Heidelberg, 2003.

8. Aditya Sai Srinivas, T., Somula Ramasubbareddy, and K. Govinda. "Loan Default Prediction Using Machine Learning Techniques." In *Innovations in Computer Science and Engineering*, pp. 529-535. Springer, Singapore, 2022.

9. Alonso Robisco, Andres, and Jose Manuel Carbo Martinez. "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction." *Financial Innovation* 8, no. 1 (2022): 1-35.

10. Ramesha, Nishanth. "Machine Learning Based Approaches to Detect Loan Defaulters." In *International Conference on Advances in Computing and Data Sciences*, pp. 336-347. Springer, Cham, 2022.

11. Yang, Xinyi. "Prediction of Credit Risk based on Logistic Regression and Random Forest technique." In *Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, pp. 531-535. 2022.

12. Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. "Loan default prediction using decision trees and random forest: A comparative study." In *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012042. IOP Publishing, 2021.

13. Sayjadah, Yashna, Ibrahim Abaker Targio Hashem, Faiz Alotaibi, and Khairl Azhar Kasmiran. "Credit card default prediction using machine learning techniques." In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), pp. 1-4. IEEE, 2018.

14. Moscatelli, Mirko, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. "Corporate default forecasting with machine learning." *Expert Systems with Applications* 161 (2020): 113567..

15. García, Salvador, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Vol. 72. Cham, Switzerland: Springer International Publishing, 2015.

16. Canbek, Gürol, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights." In *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 821-826. IEEE, 2017.